The LIMSI RT'06 Lecture Transcription System

L. Lamel, E. Bilinski, G. Adda, H. Schwenk

J.L. Gauvain, C. Barras, X. Zhu

RT06 meeting
Bethesda
May 3, 2006

Overview

- Reminder of CHIL Jan'05 system
- General improvements to speech recognizer
- Acoustic model training
- Language model training
- Concentrated development work on RT05 ihm data
- Development and evaluation results
- Some comments

Reminder: CHIL 2005 System

- Same core technology as LIMSI Broadcast News transcription system
- 35k word-list, OOV rate of 0.17% on CHIL Jan'05 test set
- Two-pass decoder: initial hypothesis generation followed by an adapted decode
- Each decoding pass generates a word lattice which is expanded with a 4-gram LM
- Initial hypothesis generation with 3-gram LM, small cross-word position-dependent, gender-independent AMs
- MLLR adaptation (2 global regression classes)
- Runtime: under 6xRT
- Manual segmentations

System Improvements

- Automatic (revised) partitioner (X. Zhu talk tomorrow)
  - Iterative GMM clustering replaced by BIC clustering
  - Additional GMM-based speaker identification clustering stage
  - Multistage system reduces speaker error by up to 50% relative to BIC alone (on Fr ench and English BN)
  - Built mdm GMMs for this eval
- Updated acoustic models
  - Additional selected far-field training data merged with ihm data
  - MAP adaptation to far-field data
  - MLLT-SAT training

System Improvements - 2

- Updated language models
  - New text normalization (case-sensitive, processing of acronyms, compound words)
  - Additional training texts (transcripts and proceedings)
  - New method for vocabulary selection
  - Neural network language model
- Revised (slightly) decoding
- Runtime: 4-10xRT

Acoustic Front-end

- PLP-like analysis
- 8kHz bandwidth
- 39 features: 12 cepstrum coefficients and the log energy, 1st and 2nd derivitives
- Cepstral mean and variance normalization (by segment cluster)

Acoustic Training Data

2005 data: Almost 97h of data from 4 sources

- **TED:** lapel mics, 39 lectures, 9.3h
- **ISL:** lapel mics, 18 meetings, 10.3h
- **ICSI:** head mounted mics, 75 meetings, 59.9h
- **NIST:** head mounted mics, 19 meetings, 17.2h
- **ICSI:** tabletop, 75 meetings 70h
- **CHIL:** head mounted, 17 seminars, 6.2h

Acoustic Models

- Triphone models with 32 Gaussian mixtures
- Separate cross-word/word-internal statistics
- Tied states with decision tree, 152 questions (position, distinctive features, neighbors)
- Gender-independent models
- MLLT, SAT
- Small models used in pass1: 5k contexts,
  5.2k tied states, 165k Gaussians
- Larger 2nd pass models: 24k contexts,
  11.5k tied states, 360k Gaussians

Language Model Training Texts

- Audio transcriptions:
  - TED: 71k words
  - NIST: 156k words
  - ISL: 116k words
  - ICSI: 785k words
- Conversational telephone speech: 3M words
- Additional transcriptions:
  - AMI/IDIAP meeting: 143k words
  - NIST RT04, RT05 data: 57k words
  - Jun04/Jan05 seminars: 55k words
  - CHIL seminars: 38k words
- BN transcriptions did not reduce perplexity

Proceedings Texts

Total: 20k articles, 42M words

TED texts:	426 papers	929k words
ASRU’99-05:	427 papers	1140k words
DARPA’97-99,04:	119 papers	317k words
Eurospeech’97-05:	3485 papers	7650k words
ICASSP’95-05:	7831 papers	14318k words
ICME’00,03:	996 papers	2101k words
ICSLP’96-04:	3202 papers	7198k words
LREC’02,04:	891 papers	2553k words
ISCA+other workshops:	2333 papers	6077k words

Text Processing

- Scripts similar to those provided by IRST
- Convert ps and pdf files to text
- Remove undesirable data: email, addresses, mathematical formulas and symbols, figures, tables, references
- Remove special formatting characters and ill-formed lines
- Stricter filtering than last year
- Transformation of acronyms and compound words (AHD, name lists)
- Case sensitive (capitalization of first word)
  - Many words have multiple forms
  - Graph with all possible caseings, use special interpolated LM
- Goal is to increase information for downstream processing

2005 Lexicon

- Wordlist optimized on dev segments from jun04 and jan05
- 20k wordlist from audio transcripts only: OOV 1.3%
- Wordlist contains all words in transcriptions and words occurring more than 3 times in proceedings texts (55k)
- Filtered by master dictionary
- 35k wordlist OOV 0.23% on jun04, 0.17% on jan05 data
- OOV rate of 0.61% on dev data RT05s eval

2006 Lexicon

- Wordlist optimized on transcripts of RT06 dev data
- 75k most probable words selected by linear interpolation (8 seminar sources + proceedings)
- CTS data not used for vocabulary selection
- Filtered by master dictionary
- 57769 words,73480 pronunciations
- OOV 0.46% on dev data (RT05s eval)
- Last year’s 35k wordlist OOV rate of 0.61%

Language Models

- Interpolation of 3 LMs trained on the text sources (weights 0.6, 0.3, 0.1)

  Seminar and meeting transcriptions (1.42M words)

  Proceedings texts (46M words)

  Transcriptions of CTS data (29M words)

- Perplexity of dev data

  4-gram: 130 (140 with last year's model)

  3-gram: 132

  2-gram: 153

Neural Network Language Model

- Recent approach that tries to attack the data sparseness problem (Bengio'01)
- Projection of the word indices onto a continuous space
- n-gram probability estimation in this continuous space

⇒ Better generalization to unseen n-grams can be expected

- Trained on audio transcripts and proceedings texts (not CTS)
- Interpolated with the back-off LM
- Efficient algorithms to train and use the neural network LM (lattice rescoring in 0.3xRT)
- Reduced perplexity on development data from 130 to 108

Decoding Strategy

- Initial hypothesis generation with 3-gram LM, small cross-word position-dependent, gender-independent AMs
- Lattice rescoring with 4-gram
- MLLR adaptation and word lattice generation (2 global regression classes) with 2-gram LM and large cross-word position-dependent, gender-independent AMs
- Lattice expansion with 4-gram LM
- Consensus decoding with pronunciation probabilities
- Rescoring with a neural network LM for the last pass

Development Data

ihm:	CHIL_20041123-0900-E[12]_h01_001
	CHIL_20041123-1000-E[12]_h01_001
	CHIL_20041123-1100-E[123]_h01_001
	CHIL_20041123-1500-E[12]_h01_001
	CHIL_20041123-1600-E[12]_h01_001
	CHIL_20041124-1000-E[12]_h01_001
	CHIL_20041124-1100-E[12]_h01_001
	CHIL_20050112-0000-E[12]_h01_001
	CHIL_20050126-0000-E1_h01_001
	CHIL_20050127-0000-E1_h01_001
	CHIL_20050128-0000-E[12]_h01_001
	CHIL_20050202-0000-E1[2]_h01_001
	CHIL_20050214-0000-E1_h01_001
	CHIL_20050310-0000-E[12]_h01_001
	CHIL_20050310-0001-E1_h01_001
	CHIL_20050314-0000-E[12]_h01_001
sdm/mdm:	CHIL_20041123-1600-E1: d01, d02, d03,d04, d05
	CHIL_20050202-0000-E2: d01, d02, d03,d04, d05
	CHIL_20050314-0000-E2: d01, d02, d03,d04, d05
	CHIL_20050128-0000-E1: d01, d02, d03,d04, d05
	CHIL_20050310-0001-E1: d01, d02, d03,d04, d05

Summary of Development Results - IHM

System	*WER (%)*
Baseline, 35k LM	26.1
Updated AM 1, 35k LM	25.9
Updated AM 2, 35k LM	25.7
Updated AM+SAT, 35k LM	25.0
58k wordlist, LM	26.0
58k LM, updated AM	25.3
+ tuning	24.6
+ SAT	24.0
+ pron probs	23.5
+ NNLM	22.6

Summary of Development Results - SDM/MDM

	sdm WER (%)		*mdm WER (%)*
System	*overlap*	*non overlap*	*overlap*
58k LM, update AM 2	64.0	62.9	
58k LM, adapt AM with FF	62.5	61.3	55.7
+ tuning	60.4	58.8	
+ SAT	60.1	58.5	
+ mdm partitioner	56.6	57.1	53.3
+ pron prob	56.1	55.3	
+ NNLM	55.2	54.4	51.9

Summary of RT06 Results - IHM

- *No cross-talk removal attempted (WER 147% due to 22k insertions)*
- Withdrew IHM result from evaluation
- Rescored with manual and UKA segmentations

System	*Cor*	*Subs*	*Del*	*Ins*	*WER*
UKA segments	72.1	18.3	9.6	22.9	50.8
ihm baseline	64.9	26.0	9.0	8.3	43.4
ihm RT06s	74.3	19.5	6.3	5.3	31.0

Summary of RT06 Results - SDM/MDM/MM3a

System	*Cor*	*Subs*	*Del*	*Ins*	*WER*
sdm					65.4
mdm	48.6	40.9	10.5	15.9	67.4
mdm (tuned Rover)	42.9	25.2	31.9	4.4	61.5
mm3a baseline	30.8	21.8	47.4	1.5	70.7
mm3a post RT06	46.1	33.9	20.0	7.5	60.1

- Retuned Rover post-eval result
- Adapted acoustic models with some of the beamformed dev data (provided by UKA)
- Baseline is 2005 system

Summary

- System development using (non-representative) data, very difficult to follow
- Acoustic model training on ihm and some farfield data but essentially no training data
- Large differences in test conditions from last year
  - Automatic partitioning
  - Seminars from multiple sites
  - More interactivity
  - Multiple microphones
- Rover combination for mdm condition
- Most techniques ported to this task